DOCUMENT RESUME

ED 449 211 TM 032 356

AUTHOR Deegear, James
TITLE Recent Literature on Whether Statistical Significance Tests Should or Should Not Be Banned.
PUB DATE 2001-02-03
NOTE 23p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, February 1-3, 2001).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; Literature Reviews; *Research Methodology; *Statistical Significance; Test Use
IDENTIFIERS Research Replication

ABSTRACT

This paper summarizes the literature regarding statistical significant testing with an emphasis on recent literature in various discipline and literature exploring why researchers have demonstrably failed to be influenced by the American Psychological Association publication manual's encouragement to report effect sizes. Also considered are defenses of statistical significance. Statistical significance testing can become a measure of the sample size a researcher has used to analyze data, in that the premium placed on obtaining statistical significance overshadows what the data actually indicate about the hypotheses. Perhaps the most productive course of action is to identify a hybrid approach that uses arguments from both sides in the debate over statistical significance testing. Reporting of statistical significance findings may continue, but researchers may be required by scholarly journals to report effect sizes and support for replication and extension studies. (Contains 30 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Recent Literature on Whether Statistical Significance Tests Should or Should Not be Banned

James Deegear

Texas A&M University 77843-4225

ED 449 211

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Deegear

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM032356

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 3, 2001.

BEST COPY AVAILABLE

Abstract

The present paper summarizes the literature regarding statistical significance testing with an emphasis on (a) recent literature in various disciplines and (b) literature exploring **why** researchers have demonstrably failed to be influenced by the APA publication manual "encouragement" to report effect sizes. Also considered are defenses of statistical significance.

Recent Literature on Whether Statistical Significance Tests Should or Should Not be Banned

Researchers have long placed a premium on the use of statistical significance testing, notwithstanding withering criticisms of many conventional practices as regards statistical inference (e.g., Carver, 1978; Meehl, 1978; Thompson, 1993, 1998a). A series of articles on these issues appeared in recent editions of the American Psychologist (e.g., Cohen, 1990; Kupfersmid, 1988; Rosnow & Rosenthal, 1989). Especially noteworthy are recent articles by Cohen (1994), Kirk (1996), Schmidt (1996), and Thompson (1996).

For example, Rozeboom (1997) recently argued that:

> Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students... [I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism... (p. 335)

And Tryon (1998) recently lamented in the American Psychologist,

> [T]he fact that statistical experts and investigators publishing the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress towards correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on scientific discipline, but the deleterious effects are doubtless substantial... (p. 796)

The older commentary eventually led to a very important change in the 1994 APA publication manual: an "encouragement" (p. 18) to always report effect sizes. Yet 11

empirical studies of either 1 or 2 volumes of 23 journals now show that this encouragement has had no effect (e.g., Kirk, 1996; Thompson & Snyder, 1998).

Anderson, Burnham and Thompson (2000) also provide yet more evidence that the "encouragement" to report effect sizes has been ineffective. And they provide a chart summarizing across both decades and diverse disciplines the frequencies of publications of criticisms of statistical significance testing.

Indeed, the recently published report of the APA Task Force on Statistical Inference says in two different locations that effect size should always be reported for all primary results (Wilkinson & APA Task Force on Statistical Inference, 1999). Yet the Task Force (1999) itself acknowledged that "Unfortunately, empirical studies of various journals indicate that the effect size of this [APA publication manual] encouragement has been negligible" (p. 599).

The present paper explores these views in detail with an emphasis on (a) important recent literature in various disciplines (cf. Finch, Cumming & Thompson, 2001; Hubbard & Ryan, 2000) and (b) literature exploring **why** researchers have failed to be influenced by the publication manual "encouragement" (cf. Thompson, 1998b). Also considered are defenses of statistical significance (cf. Hagen, 1998; Kover, 2000; Mick, 2000; Robinson & Levin, 1997; Stewart, 2000); some of these treatments have been thoughtful, but others have been seriously flawed (see Thompson, 1998a).

Notwithstanding the movement of the field away from overemphasis on statistical significance, it remains important to understand the flawed logic of those abusing statistical tests. As Thompson (1996) noted:

> We must understand the bad implicit logic of persons who misuse statistical tests if we are to have any hope of persuading them to alter their practices – it will not be sufficient merely to tell researchers not to use statistical tests, or to use them more judiciously. (p. 26)

Historical Growth and Continued Use of Significance Test

Hubbard and Ryan (2000) examined 8,001 empirical articles in 12 APA journals for use of statistical significance testing (SST) from 1911 to 1998. They found that 6,589 (82.4%) of the articles employed SST. Further, regression analysis examining the growth of the use of SST over time from 1911 to 1998 indicated an R^2=.945 (R=.972). That is, during more than three-quarters of a century we have witnessed a consistent, expansive use of SST. In face of the aforementioned criticisms against SST, why has this happened?

Hubbard and Ryan (2000) asserted that an "inference revolution" in psychology from 1940 to 1955 involved the adoption of a hybrid methodology of Fisherian and Neyman-Pearson statistical concepts. Cohen (1990) asserted that Fisher's ideas of inductive inference became the standard in behavioral sciences because "they were very attractive. They offered a deterministic scheme, mechanical and objective, independent of content, and led to clear-cut yes-no decisions" (p. 1307).

Hubbard and Ryan (2000) contended that, subsequent to Fisher's introduction of his statistical approaches, Nyeman and Pearson's concepts of hypothesis testing, alternative hypothesis, Type I and II errors, and statistical power were conceived. These were then combined with Fisher's statistical methodologies (Shaver, 1993) to form the present practice of empirical research focused on rejection of the null hypothesis at some

p (usually .01 or .05). Consequently, Hubbard and Ryan, in their review of 12 journals, demonstrated the dramatic increase in the implementation of SST. The percentage of empirical articles using SST from 1911 to 1929 averaged 17%. This average increased to 85% by 1960 and grew beyond 90% after 1970. Such dramatic increases must be the results of some perceived utility in SST.

Arguments in Favor of Significance Testing

From the above discussion, it is evident that SST continues to enjoy wide-ranging usage. Surely, researchers would not use a faulty methodology for the advancement of science? *If* that is the case, then what are the reasons for its support?

Contention that SST *is* Logical

Hagen (1998) has been one of the more ardent supporters of SST. His arguments and rebuttals against critics center on what he considers to be the logic of SST. Criticism of SST has included the assertion that the null hypothesis is a statement about the sample, not the population from which the sample is drawn (Thompson, 1998a). But what we **want** to do is use the sample to make an inference to the population, even though SST does **not** do this (Cohen, 1994).

Thompson further asserts that the "nil" null hypothesis (Cohen, 1994) is always false. Therefore, the "nil" null will always be rejected at some sample size (Thompson & Kieffer, 2000). Furthermore, what would be the logic of investigating that which is already known? SST would then be nothing more than a search for what is already known. In response, Hagen asserted that the null hypothesis is not a statement about the sample. Instead, he contends that it is a statement about the population; "The null hypothesis is a statement about the population from which the sample is drawn" (p. 801).

Therefore, Hagen's logic implies belief that SST results from a sample may be extended to the population.

Second, Hagen (1998), in response to Thompson's (1998a) argument that populations do not exist where the nil null hypothesis is exactly true, claimed that "If the null hypothesis is always false, then everything would have to be related to everything else" (p. 801). That is, if there always exist some degree of relationship (via extension of the argument that null hypotheses are always false), then all measurable traits would have to at least some degree be related to any cosmic contention one would like to make. Hagen appears to consider such contentions fallible. Hagen suggested that because it therefore can *not* be claimed that the null hypothesis is always false, there exists a cause for investigation by SST methods. Of course, many believe that the "nil" null is always false (Cohen, 1994; Thompson, 1996).

Finally, Hagen (1998) asserted that the logic underlying SST is valid. He contends that SST relies on what he calls "proof by contradiction" (p. 802). Through analogy of a criminal investigation, Hagen stated that sound reasoning supports the use of SST. He further claimed that if such reasoning is accepted in arenas outside of scientific inquiries (i.e., courtrooms), then "scientists are on shaky ground if they deny its usefulness in the lab" (p. 802).

The Positive Reinforcement of SST

Mick (2000) claimed that marketing researchers are proponents of SST. He stated that in marketing research it is of utmost importance to be able to distinguish between findings that are "due to chance" and those that are not. The necessity of being able to make such distinctions is extended to academia's continued reinforcement of SST

methodologies. The function of research (and, it would seem, marketing), he stated, is to challenge previously held beliefs. As such, SST and subsequent claims that one has made a "significant" finding are the primary means by which to challenge beliefs and to direct attention to one's findings and, more importantly, to one's self. It is through such notoriety (Kover, 2000; Mick, 2000; Rosnow & Rosenthal, 1989) that graduate students and faculty rise in the ranks of prestige (not to mention paycheck size and employment security). Replicating the research findings of others' groundbreaking revelations that have led to their respective notoriety and advancement does not attract attention to one's self. The system of acknowledgment and accolade is set up to reward challenges and reversals of previously held dogma, not to support replication and extension of previous work.

SST is the Best Alternative

Proponents of SST assert that there do not exist viable alternatives (i.e., Stewart, 2000; Winer, 2000). Stewart (2000) is one of the more recent advancers of this position. Simply put, he claimed that without the guidance of SST, we would not be able to discern what is from what is not important in research findings. He stated that SST serves as a "useful screen" (p. 686) for attention-getting as well as serving as a standard for editors and reviewers for making decisions about what is or is not to be published.

Although Stewart (2000) asserted that effect size reporting is useful and should be required, he does not see it as a replacement for SST. He claimed that it is "fraught with far more problems than statistical significance testing" (p. 687). Stewart claimed that suggestions for categorization of effect sizes do not help to delineate what findings are substantive, important research. The presence of large effect sizes in marketing research

does not always attract attention. Indeed, in marketing, Stewart pointed out, large effect sizes are relatively uninteresting. It is the small effect sizes associated with marketing and advertising (those over which the marketer has some control) that are theoretically and practically important. Thus, while effect sizes may be important, they should not be the determining factor to what we should attend.

Third, Stewart (2000) asserted that although replication and meta-analysis research are important tools in the development of a cumulative knowledge base and the elimination of alternative hypotheses, they should not replace SST. On this point, Stewart appears to agree with the aforementioned contentions (Kover, 2000; Mick, 2000; Rosnow & Rosenthal, 1989) that replication studies do not carry the weight of SST and are therefore considered less interesting. He also argued that exact replications of studies are difficult to achieve. Finally, he questioned what constitutes a replication, especially in the absence of SST. What, he asks, should be considered successful replications: findings in the same direction of previous research or should similar effect sizes be required for successful replication? Consequently, he would assert that SST remains the only viable, objective guideline for determining importance of findings.

SST is not Flawed, Researchers are Flawed

Many of the proponents of SST acknowledge that SST has its limitations and can benefit from complementary approaches (Hagen, 1998; Stewart, 2000). However, they assert that it is often the researchers' misuses and misinterpretations of SST that is flawed, not the methodology itself. Stewart commented that,

> As an editor, I am quite agnostic with respect to statistical significance testing; it is the way the test is used and presence or absence of additional information and

analyses that determine the appropriateness of a test and any conclusions that are derived from such a test. (p. 689)

Hubbard and Ryan (2000) outlined many of the arguments for the continued use of SST. According to their article, although supporters of SST recognized its abuses and misuses, they cautioned that doing away with SST would be analogous to "throwing the baby out with the bath water" (p. 675). The assertion is not to get rid of a useful tool just because the user misapplies it. Rather, they implore users to apply the tool "judiciously" and complement it with other information. Indeed, Hagen (1998), commented, "I am struck by the beauty, elegance, and usefulness of NHST [null hypothesis significance testing], but other methods of inference may be equally elegant and even more useful depending on the question being asked" (p. 803).

Some Arguments Against Statistical Significance Testing

Arguments against the usefulness of SST extend back to the early 1900's (see Boring, 1919). Hubbard and Ryan (2000) cited numerous criticisms of SST through out the past 4 decades. In fact, Tryon (1998) cited research indicating that criticism of null hypothesis significance testing began immediately after Fisher first introduced it in 1925. The full litany of reasoning against the use of SST is beyond the scope of this paper. But some of the key arguments this writer identified are examined below.

Misinterpretations of SST

The presence of a "statistically significant" finding has often resulted in erroneous and unjustified interpretations and claims. Tryon (1998) lamented that, "the fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing" (p. 796).

What exactly does one mean by "statistically significant"? To answer this, we first must understand what statistical significance asks. Thompson (1994) stated that statistical significance addresses the following question:

> Assuming the sample data came from a population in which the null hypothesis is (exactly) true, and given our sample statistics and sample size(s), is the calculated probability of our sample results less than the acceptable limit ($p_{critical}$) imposed regarding a Type I error? (p. 2)

Put another way, Thompson (1994) stated the question as, "What is the probability of the sample statistics?" To state that statistical significance has been achieved at some *p* value (i.e., $p_{critical} > p_{calculated}$) is to be able to reject the null hypothesis. Consequently, this decision means "that we believe our sample results are relatively unlikely, given our assumptions, including our assumption that the null hypothesis is exactly true" (Thompson, 1994, p. 3) in the population.

The problem is that many researchers then extrapolate findings from the sample to the population. They reason that because a sample comes from a given population, any property of that sample must then exist in the population. In reality, the inference is from the population to the sample. As Thompson (1996) noted, "given an assumption about the parameters 'B' [of a population], what is the likelihood of 'A,' the sample statics?" (p. 28). Boring (1919) long ago noticed this problem in significance testing: "A knowledge of the 'probability that a difference is not due to chance' is distinctly worthwhile on the descriptive side; but this measure of significance does not necessarily apply to the general class for which a sample stands" (p. 337). Generalizations from a

sample to a population should only be made through repeated replications of the sample statistics.

Moreover, Kirk (1996) contended that "*even when* a significance test is interpreted correctly, the business of science does not progress" (p. 753, emphasis added). Kirk's point was that despite correct usage, SST alone is not adequate for sound scientific advancement. Findings or absence of significance, he argued, are not the endgame; magnitudes of differences are what is important. He continued:

> The appeal of null hypothesis significance testing is that it is considered to be an objective, scientific procedure for advancing knowledge. In fact, focusing on *p* values and rejecting null hypotheses actually distracts us from our real goals: deciding whether data support our scientific hypothesis and are practically significant or useful. (p. 755)

The Forced-Choice Problem

The use of p-values for determining whether or not the results of a statistical test are statistically significant forces researchers into an arbitrary decision about what should and should not be considered significant (which is then often mistaken for what is an important finding). Cohen (1990) asserted that the yes-no decision for determining statistical significance was appropriate in the field of agronomy (from which Fisher hailed) where outcome research directly affected agrarian decision-making. But, as Cohen aptly pointed out, "we do not deal in manure, at least not knowingly" (p. 1307).

Rosnow and Rosenthal (1989) reminded us that Fisher also objected to the use of a fixed point for dichotomous decision making. By arbitrarily creating a point at which results are either statistically significant or not significant, many important results are not

pursued, not published, or ignored. As Cohen (1990) asserted, research is not designed to resolve issues, but rather to indicate increased or decreased likelihood of a given position. Cohen further argued that any given p-value "is not a cliff but a convenient reference point along the possibility-probability continuum" (p. 1311).

Carver (1993) offered a further argument against the driving force of p-values. He asserted that results should be interpreted first with respect to the data and second with respect to statistical significance. Doing so places the cart behind the horse. Researchers are then forced to consider their data as regards their hypotheses instead of ascribing finality of statistical significance or nonsignificance and then moving on to new research. Kirk (1996) stated, "focusing on *p* values and rejecting null hypotheses actually distracts us from our real goals: deciding whether data support our scientific hypothesis and are *practically* significant or useful" (p. 775, emphasis added). Finaliy, and succinctly, Rosnow and Rosenthal (1989) offered perhaps the most eloquently stated opposition to a fixed *p* value for determining whether or not results are significant: "That is, we want to underscore that, surely, God loves the .06 nearly as much as the .05" (p. 1277).

A Matter of N

Statistical Significance testing often becomes a measure of the sample size a researcher has employed to analyze data. The premium placed on obtaining statistical significance overshadows what the data actually indicate about the hypotheses. Yet, obtaining statistical significance appears not to be a function of the obtained data, but a function of the number of participants enlisted in the study. As Thompson (1993) pointed out, because the null hypothesis is never exactly true, at some sample size

statistical significance will always be achieved (Cohen, 1990; Thompson, 1993; Thompson, 1998a). Therefore, the researcher need only obtain sufficient numbers of participants to obtain this difference. As Thompson (1998a) stated, "statistical testing becomes a tautological search for enough participants to achieve statistical significance. If we fail to reject [the null hypothesis], it is only because we've been too lazy to drag in enough participants" (p. 799).

Why Significance Testing Continues in Such Prominence

Despite widespread criticism of the fallacies of SST, it continues to be overwhelmingly used (e.g., Hubbard & Ryan, 2000; Kirk, 1996). The American Psychological Association has attempted to steer researchers away from reliance on SST by encouraging authors to report effect sizes in articles they submit for publication. The Publication Manual of the American Psychological Association (4th ed.) (1994) stated:

> Neither of the two types of probability values reflects the importance (magnitude) of an effect or the strength of a relationship because both probability values depend on sample size.... You are encouraged to provide effect-size information. (p. 18)

Although this would appear to be a step in the right direction, this statement lacks the directive many researchers appear to need. Thompson (1998b) reported that the APA's encouragement has had little appreciable effect upon researchers' publication habits. Thompson's suggestion as to the absence of effect is as follows:

> Merely encouraging effect-size reporting is akin to invoking Santa Claus or the Tooth Fairy as the sole incentive for reasonable behavior: Some kids are more susceptible than others to these influences, the incentive's vagueness may itself

> help render some temptations too strong for even the nicest kids to resist, and the incentive system may not be operational on a daily basis throughout the year. More important, to only encourage what is minimally reasonable behavior creates an inherent contradiction and thereby inevitably weakens the moral authority of the admonition. (p. 337)

Hubbard and Ryan (2000) cited a litany of reasons as to why SST persists. These include: (1) the historical force behind its use, (2) lack of understanding of the appropriate uses of SST, (3) lack of a willingness to ban its use, (4) the arguments of those who champion its use, and (5) absence of effort to de-legitimatize the use of SST.

First, Hubbard and Ryan (2000) asserted that the more than 60 years of SST usage has created an "inertia" that will be difficult to overcome. Second, despite awareness of SST's limitations, informed researchers "must occasionally bow down before" it in order to achieve publication (pp. 672-673). Third, despite overwhelming criticism against its usage, journal editors have yet to adopt policies against the use of SST. Fourth, arguments continue to persist in favor of SST. Hubbard and Ryan reported that supporters claim no superior alternatives to SST; the persistent insistence on sensible usage pressures researchers not to abandon its usage all together. Finally, Hubbard and Ryan lamented the fact that SST continues to receive primary consideration in statistics classes, thereby perpetuating its usage and misinterpretations. Consequently, "this resistance is fueled primarily by the belief that the absence of these tests in research reports would prejudice the publication efforts of their graduate students. Faculty opposition to the idea of abolishing SST is to be expected" (p. 675).

Carver (1978) asserted that one reason SST persists is that researchers mistake it as evidence of replicability. Carver stated that researchers errantly deduce that if results are statistically significant, they will replicate and if they are not statistically significant, they will not replicate. However, this is not the correct interpretation of SST. Again, statistical significance testing "is the probability (0 to 1.0) of the sample statistics, given the sample size, and assuming the sample was derived from a population in which the null hypothesis (H_o) is exactly true" (Thompson, 1996, p. 27). Notice, there is no mention of a prediction that said results may or may not occur in the future. The best way to make such an assertion is to replicate results, not to treat present results as tea leaves with predictive power.

Carver (1978) further contended that researchers rely on SST as an objective standard for identifying importance. The finding or absence of "significance" has become a convenient means for interpreting a study. Carver stated, "Statistical significance testing has purportedly provided an objective, although inappropriate, solution to the problem of deciding whether a result is important" (p. 393).

However, SST has been shown to be anything but objective. Rosnow and Rosenthal (1989) stated, "determining the particular level of significance of the data at which a null hypothesis will be rejected is essentially a personal decision" (p. 1277). Cohen (1990) appeared to support this contention: "The prevailing yes-no decision at the magic .05 level from a single research is a far cry from the use of informed judgment" (p. 1311).

Conclusion

This paper has explored the ongoing debate over the utility of statistical significance testing. Arguments for and against SST have been examined. It would appear that the debate will continue for some time to come. Perhaps the most prudent course of action is to identify a hybrid approach utilizing arguments from both sides. For example, while reporting of SST findings may continue, researchers may be mandated to report effect sizes and supported for replication and extension studies. Indeed, editors at various journals have now adopted editorial policies "requiring" that effect sizes be reported. These include:

--Career Development Quarterly

--Contemporary Educational Psychology

--Educational and Psychological Measurement

--Journal of Agricultural Education

--Journal of Applied Psychology

--Journal of Consulting and Clinical Psychology

--Journal of Early Intervention

--Journal of Experimental Education

--Journal of Learning Disabilities

--Language Learning

--Measurement and Evaluation in Counseling and Development

--The Professional Educator

--Research in the Schools.

It is evident that SST will not soon be abandoned. However, the field is clearly moving away from reliance solely on statistical significance tests to evaluate the noteworthiness of research results (Kirk, 2001); Shibley Hyde, 2001; Vacha-Haase, 2001). Thompson (in press) reviews some of these alternative choices.

References

American Psychological Association. (1994). Publication Manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Anderson, D.R., Burnham, K.P., & Thomson, W. (2000). Null hypothesis testing: Problems, prevelance, and an alternative. Journal of Wildlife Management, 64, 912-923.

Boring, E.G. (1919). Mathematical vs. scientific importance. Psychological Bulletin, 16, 335-338.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61(4), 287-292.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Finch, S., Cumming, G., & Thompson, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. Educational and Psychological Measurement.

Hagen, R.L. (1998). A further look at wrong reasons to abandon statistical testing. American Psychologist, 53, 801-803.

Hubbard, R., & Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology – and its future prospects. Educational and Psychological Measurement, 60, 661-681.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Kirk, R.E. (2001). Promoting good statistical practices: Some suggestions. Educational and Psychological Measurement, 61(2).

Kover, A.J. (2000). A response to the Hubbard and Ryan article. Educational and Psychological Measurement, 60, 691-692.

Kupersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Mick, D.G. (2000). In search of significance (statistical and otherwise). Educational and Psychological Measurement, 60,682-684.

Robinson, D.H., & Levin, J.R. (1997). Reflections on statistical and substantive significance, with a slice of replication. Educational Researcher, 26, 21-26.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Rozeboom, W.W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 335-392). Mahwah, NJ: Erlbaum.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1, 115-129.

Shaver, J.P. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61(4), 293-316.

Shibley Hyde, J. (2001). Reporting effect sizes: The roles of editors, textbook authors, and publication manuals. Educational and Psychological Measurement, 61(2).

Stewart, D.W. (2000). Testing statistical significance testing: some observations of an agnostic. Educational and Psychological Measurement, 60, 685-690.

Thompson, B. (1993). Theme issue: Statistical significance testing in contemporary practice. Journal of Experimental Education, 61(4).

Thompson, B. (1994). The concept of statistical significance testing (An ERIC/AE Clearinghouse Digest #EDO-TM-94-1). Measurement Update, 4(1), 5-6. (ERIC Document Reproduction Service No. ED 366 654)

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Thompson, B. (1998a). In praise of brilliance: Where that praise really belongs. American Psychologist, 53, 799-800.

Thompson, B. (1998b). Review of What if there were no significance tests? By L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.). Educational and Psychological Measurement, 58, 165-181.

Thompson, B. (in press). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? Journal of Counseling and Development.

Thompson, B., & Snyder, P.A. (1998). Statistical significance and reliability analyses in recent JCD research articles. Journal of Counseling and Development, 76, 436-441.

Tryon, W.W. (1998). The inscrutable null hypothesis. American Psychologist, 53, 796.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604.

Winer, R.S. (2000). Comments on "the historical growth of statistical significance testing in psychology – and its future prospects". Educational and Psychological Measurement, 60, 693-696.